# Semi-Supervised Cause Identification from Aviation Safety Reports

Isaac Persing and Vincent Ng
University of Texas at Dallas

## Background

- The **Aviation Safety Reporting System** (**ASRS**) collects voluntarily-submitted reports on aviation safety incidents.
  - Each report includes a narrative describing an incident.
- **Cause Identification** is the task of identifying all and only those causes (or shaping factors) that contributed to each incident given a narrative describing it.
- 14 causes (or **shapers**)
  - **Attitude**, **Communication Environment**, **Duty Cycle**, **Familiarity**, **Illusion**, **Pressure**, **Physical Environment**, **Physical Factors**, **Preoccupation**, **Proficiency**, **Resource Deficiency**, **Taskload**, **Unexpected**, and **Other**.
  - Each incident may be caused by one or more of these factors.

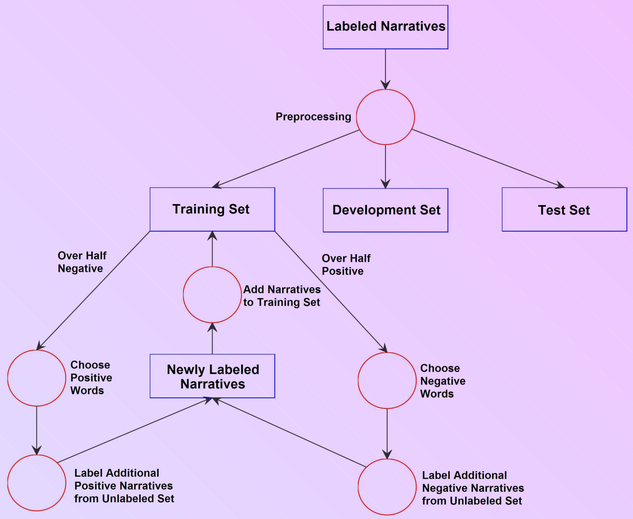
## Challenges

- **Multi-class classifiation**
  - While many classification problems involve only a few classes, Cause Identification involves 14 classes, one for each shaper.
- **Skewed class distributions**
  - Some shapers contribute to as few as 0.2% of incidents (Illusion), or to as many as 48.9% of incidents (Resource Deficiency).
- **Multi-label categorization**
  - Incidents may be caused by multiple shapers.
- **Presence of irrelevant information**
  - Most of the information in a narrative (e.g., where and when the incident took place, who was around when the incident took place, opinions of other people involved) is not pertinent to Cause Identification.
- **Scarcity of labeled data**
  - Until we worked on this problem, there was no publicly-available corpus in which the reports are labeled with shapers. For training and evaluation, we hand-annotated a small dataset consisting of **1,333 narratives**.

## Dealing with Challenges

- We treat Cause Identification as **14 binary classification tasks**. Thus each report may be labeled as a positive example of multiple shapers by 14 different SVM classifiers.
- Since some shapers appear only rarely, we identify the 10 least frequently occurring shapers as **minorities** and treat **minority shaper classification** as a separate problem.
  - Cumulatively, these 10 minority shapers account for only 26.2% of labels.
  - The minority shapers are: **Attitude**, **Communication Environment**, **Duty Cycle**, **Familiarity**, **Illusion**, **Pressure**, **Physical Factors**, **Preoccupation**, **Taskload**, and **Unexpected**.
- ASRS archives many narratives that have not been annotated with shaper information. We use a bootstrapping algorithm to automatically label additional reports from the large remaining unlabeled set. This helps us overcome the small human annotated data set size.

## Bootstrapping Algorithm




**Input**: L+ (a small set of positively-labeled narratives)
L- (a small set of negatively-labeled narratives)
U (a large set of unlabeled narratives)

1. Preprocess the documents (e.g., acronym expansion, stemming)

1. For each Shaper:
   1. Repeat until done:
      1. If L- is larger than L+,
         1. Choose 4 words that are highly positively correlated with the narratives in L+ using log likelihood ratios.
         2. Add to L+ any narrative in U having at least 3 positive words.
      2. Else
         1. Choose 4 words that are highly positively correlated with the narratives in L- using log likelihood ratios.
         2. Add to L- any narrative in U having at least 3 negative words.
   2. Train an SVM classifier on all documents in L+ and L-.

**Expand L+** / **Expand L-**

## Sample Word Selections

| Shaper | Positive Expanders | Negative Expanders |
|---|---|---|
| **Familiarity** | unfamiliar, layout, unfamiliarity, rely | |
| **Physical Environment** | cloud, snow, ice, wind | |
| **Physical Factors** | fatigue, tire, night, rest, hotel, awake, sleep, sick | declare, emergency, advisory, separation |
| **Preoccupation** | distract, preoccupied, awareness, situational, task | declare, ice, snow, crash, fire, rescue, anti, smoke |

## Experimental Setup

- Goal: enhance the performance of a Cause Identification system by making use of **unlabeled data**, as opposed to a purely supervised system which would train only on human-annotated data.
- 5-fold cross-validation with the 1,333 human-annotated reports
- Each SVM classifier may attempt to tune zero, one, or both of two parameters
  - **Classification Threshold** - How confident about the classification must one of the SVMs be before we label the narrative as a positive example of the SVM's shaper? Default value is 0.5.
    - Our goal is to maximize **F-measure** on the Cause Identification task. Classification Threshold helps us find the right balance between precision and recall to do this.
  - **Iteration** - How many iterations of the bootstrapping algorithm should we perform before training the SVMs on the resulting training data? Default value is 0.
    - On each iteration, we might add some noise to the training data. The iteration parameter tells us when to stop bootstrapping.
  - Systems tuning any of these parameters train on 3/5 of the initially labeled data, use 1/5 for parameter tuning, and 1/5 for testing.
  - The system that uses default values for all these parameters trains on 4/5 of the initially labeled data and tests on the remaining 1/5.
- We trained 2 purely **supervised** SVM baseline systems.
  - **$B_{0.5}$** does no parameter tuning.
  - **$B_{ct}$** tunes the classification threshold parameter, but not the iteration parameter.
  - Both use **only human-annotated data** for training.
- We trained 2 **semi-supervised** SVM systems using our bootstrapping algorithm.
  - **$E_{0.5}$** tunes the iteration parameter, but not the classification threshold parameter.
  - **$E_{ct}$** tunes both classification threshold and iteration parameters.
  - Both use **human-annotated and unlabeled data** for training.

## Results and Discussion

| System | All 14 Classes | | | 10 Minority Classes | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Precision | Recall | F-measure |
| **$B_{0.5}$** | 67.0 | 34.4 | 45.4 | 68.3 | 23.9 | 35.4 |
| **$B_{ct}$** | 47.4 | 59.2 | 52.7 | 47.8 | 34.3 | 39.9 |
| **$E_{0.5}$** | 60.9 | 40.4 | 48.6 | 53.2 | 35.3 | 42.4 |
| **$E_{ct}$** | 50.5 | 54.9 | 52.6 | 49.1 | 39.4 | 43.7 |

- Neither SVM classifier with bootstrapping outperforms the baseline $B_{ct}$ on the 14 shaper classification task.
- Both SVM classifiers with bootstrapping outperform both baselines on the 10 minority shaper classification task.
  - In particular, $E_{ct}$ obtains a micro f-measure of 43.7%, a relative error reduction of 6.3% over $B_{ct}$'s performance (39.9%)
- Bootstrapping is useful when there are few positive examples of a class, but when more positive examples are available, noise introduced hurts performance more than the training set size increase helps.
- **Bootstrapping helps minority class prediction.**